Data Quality Analysis Results - 12,000 Reviews from 20 Apps

Key Findings

1. Dataset Overview

- 12,000 reviews from 20 popular apps (600 each - evenly distributed)
- 11,407 unique authors (95% unique reviewers)
- All 11 data fields present

E.g. [


```
  {
    "review_id": "45ec1dd0-2dde-4a0f-a362-48d56b93bb42",
    "app_id": "com.whatsapp",
    "author": "John Doe",
    "rating": 5,
    "content": "Great app! Love the new features.",
    "timestamp": "2026-01-28T15:47:37",
    "thumbs_up": 12,
    "app_version": "2.26.2.72",
    "reply_content": "Thanks for your feedback!",
    "reply_timestamp": "2026-01-29T10:00:00",
    "scraped_at": "2026-01-29T15:47:50.123456"
  }
]
```


2. Missing Values

Field	Missing	Impact
app_version	14.2%	Minor - not critical for sentiment
reply_content	86.3%	Expected - most reviews don't get dev replies
Core fields	0%	All essential fields complete

3. Rating Distribution - Positive Skew

- Mean: 3.71, Median: 5.0
- 58.6% are 5-star, 24.7% are 1-star (polarized)
- This J-shaped distribution is typical for app reviews

4. Text Quality Issues

Issue	Count	Percentage
Single-word reviews	2,694	22.4%
Very short (1-10 chars)	4,103	34.2%
Emoji-only	392	3.3%
All caps	89	0.7%

5. Temporal Coverage

- 14-day window (Jan 14-28, 2026)
- All recent reviews - suitable for current sentiment analysis

6. Language

- 96.6% Latin script (English-dominant)
- 15.2% contain non-ASCII characters (emojis, accents, other languages)
- No encoding errors detected

7. Developer Replies

- 13.7% of reviews have developer replies
- Higher reply rate for negative reviews (23.1% for 1-star vs 8.8% for 5-star)

Implications for Downstream Modeling

1. **Class imbalance**: The 5-star dominance (58.6%) will need handling in classification tasks
2. **Low-signal reviews**: 22.4% single-word + 34.2% very short reviews may need filtering or special handling
3. **Deduplication**: Common phrases like "good"/"nice" appear hundreds of times - consider weighting strategies
4. **Multilingual content**: 15.2% non-ASCII may require language detection/filtering for English-only models